

Michael Stephen Saxon
saxon@ucsb.edu <https://saxon.me/>

I currently work on semantic assessment and analysis of language & text-to-image models. I am an **NSF Fellow** and 3rd-year Ph.D. candidate, with 7 first-author papers in NLP and ML venues. I am **fluent in PyTorch & HuggingFace**, and have done **5 research internships** in generative text and language understanding including at **Amazon, Meta**, and a clinical startup. *Bibliometrics: $c \approx 150$; $h = 8$; $i_{10} = 7$.*

Education

University of California, Santa Barbara Santa Barbara, CA
Ph.D., Computer Science: **4.0/4.0** 9/2020–6/2025
Thesis Topic—Analyzing Semantic Capabilities in Large Generative Pretrained Models
Advisors: William Yang Wang, Ph.D.

Arizona State University Tempe, AZ
MS., Computer Engineering: **3.9/4.0** 8/2018–5/2020
Advisors: Visar Berisha, Ph.D. & Sethuraman Panchanathan, Ph.D.

Arizona State University Tempe, AZ
BSE., Electrical Engineering; *Minor*, Mathematics: **Magna Cum Laude** 8/2014–8/2018

Publications



Archival

○ **Lead Mentor** ^ **Representative** ☆ **Award**

- [p20] X. Wang, W. Zhu, **M. Saxon**, M. Steyvers, WY. Wang, “Large Language Models Are Implicitly Topic Models: Explaining and Finding Good Demonstrations for In-Context Learning,” **NeurIPS 2023**, [arXiv:2301.11916](#), May 2023
- [p23] L. Pan, **M. Saxon**, W. Xu, D. Nathani, X. Wang, WY. Wang, “Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies,” *Preprint* [arXiv:2308.03188](#), Aug 2023.
- [c22] **M. Saxon**, WY. Wang, “Multilingual Conceptual Coverage in Text-to-Image Models,” **ACL 2023**; ^ **FAccT 2023 Oral** [arXiv:2306.01735](#), [[oral presentation link](#)] Jul 2023.
- [c21] Y. Tuan, A. Albalak, W. Xu, **M. Saxon**, C. Pryor, L. Getoor, WY. Wang, “CausalDialogue: Modeling Utterance-level Causality in Conversations,” **ACL 2023 F** [arXiv:2212.10515](#), Jul 2023.
- [p19] V. Himakunthala*, A. Ouyang*, D. Rose*, R. He*, A. Mei, Y. Lu, C. Sonar, **M. Saxon**, WY. Wang, “Let’s Think Frame by Frame: Evaluating Video Chain of Thought with Video Infilling and Prediction,” *preprint*, [arXiv:2305.13903](#), May 2023
- [p18] D. Rose*, V. Himakunthala*, A. Ouyang*, R. He*, A. Mei, Y. Lu, **M. Saxon**, C. Sonar, D. Mirza, WY. Wang, “Visual Chain of Thought: Bridging Logical Gaps with Multimodal Infillings,” *preprint*, [arXiv:2305.02317](#), May 2023.
- [p17] **M. Saxon***, A. Mei*, S. Chang, ZC. Lipton, WY. Wang, “Users are the North Star for AI Transparency,” *preprint*, [arXiv:2303.05500](#), Mar 2023.
- [c16] **M. Saxon**, X. Wang, W. Xu, WY. Wang, “PECO: Examining Single Sentence Label Leakage in Natural Language Inference Datasets,” **EACL 2023** [arXiv:2112.09237](#), May 2023.
- [c15] M. Ho*, A. Sharma*, J. Chang*, **M. Saxon**, S. Levy, Y. Lu, WY. Wang, “WikiWhy: Answering and Explaining Cause-and-Effect Questions,” **ICLR 2023** *Oral (top 5%)* [arXiv:2210.12152](#), May 2023.

- [c14] X. Wang, **M. Saxon**, J. Li, H. Zhang, K. Zhang, WY. Wang, “Causal Balancing for Domain Generalization,” **ICLR 2023** [arXiv:2206.05263](#), May 2023.
- [c13] W. Xu, Y. Tuan, Y. Lu, **M. Saxon**, L. Li, WY. Wang, “Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis,” **EMNLP 2022 F** [arXiv:2210.05035](#), Dec 2022.
- [c12] W. Xu, **M. Saxon**, M. Sra, WY. Wang, “Self-Supervised Knowledge Assimilation for Expert-Layman Style Transfer,” **AAAI 2022** [arXiv:2110.02950](#), Jan 2022.
- [c11] X. Wang, W. Chen, **M. Saxon**, WY. Wang, “Counterfactual Maximum Likelihood Estimation for Training Deep Networks,” **NeurIPS 2021** [arXiv:2106.03831](#), Dec 2021.
- [c10] **M. Saxon**, S. Levy, X. Wang, A. Albalak, WY. Wang, “Modeling Disclosive Transparency in NLP Application Descriptions,” **EMNLP 2021** *Oral (8% of subs.)* [arXiv:2101.00433](#), pp. 2023–2037. 
- [c9] **M. Saxon**, S. Choudhary, J. McKenna, A. Mouchtaris, “End-to-End Spoken Language Understanding for Generalized Voice Assistants,” **Interspeech 2021**, pp. 4738–4742. 
- [c8] S. Levy, **M. Saxon**, WY. Wang, “The Truth is Out There: Investigating Conspiracy Theories in Text Generation,” [arXiv:2101.00379](#), **Findings of ACL 2021**, pp. 4718–4729.
- [j7] **M. Saxon**, A. Tripathi, Y. Jiao, J. Liss, V. Berisha, “Robust Estimation of Hypernasality in Dysarthria,” **IEEE Trans. on Audio, Speech, and Language Processing** 2020, Vol. 28, pp. 2511–2522.
- [c6] **M. Saxon***, J. McKenna*, S. Choudhary*, G. Strimel, A. Mouchtaris, “Semantic Complexity in End-to-End Spoken Language Understanding,” **Interspeech 2020**, pp. 4273–4277.
- [c5] M. Moore, P. Papreja, **M. Saxon**, V. Berisha, S. Panchanathan, “UncommonVoice: A Crowdsourced Dataset of Dysphonic Speech,” **Interspeech 2020**, pp. 2532–2536.
- [c4] M. Moore, **M. Saxon**, H. Venkateswara, V. Berisha, S. Panchanathan, “Say what? A dataset for exploring the error patterns that two ASR engines make,” **Interspeech 2019**, pp. 2528–2532.
- [c3] **M. Saxon**, J. Liss, V. Berisha, “Objective Measures of Plosive Nasalization in Hypernasal Speech,” 2019 **IEEE ICASSP 2019**, pp. 6520–6524.
- [w2] **M. Saxon***, S. Bhandari*, L. Ruskin, G. Honda, “Word Pair Convolutional Model for Happy Moment Classification,” **2nd Workshop on Affective Content Analysis, AAAI 2019**, pp. 111–119.  *(Workshop Oral; CL-Aff Shared task runner up, 2/47)*
- [c1] T. Houghton, **M. Saxon**, Z. Song, H. Nyugen, H. Jiang and H. Yu, “2D Grating Pitch Mapping of a through Silicon Via (TSV) and Solder Ball Interconnect Region Using Laser Diffraction” **IEEE 66th Electronic Components and Technology Conference (ECTC) 2016**, pp. 2222–2227.  *(Texas Instruments Best Student Interactive Paper Award)*

Non-archival Presentations

- [n6] A. Tanna, **M. Saxon**, A. El Abbadi, WY. Wang, “Data Augmentation for Diverse Voice Conversion in Noisy Environments,” **Interspeech 2023 Show and Tell** [arXiv:2305.10684](#), Aug 2023. 
- [n5] **M. Saxon**, WY. Wang, “Multilingual Conceptual Coverage in Text-to-Image Models,” **FAccT 2023**  *Oral (Non-archival)* [OpenReview:5H2m3tCEaQ](#), Jun 2023.
- [n4] **M. Saxon**, X. Wang, W. Xu, WY. Wang, “Automated Cheating Feature Semantic Identification for NLI Datasets,” **SoCal NLP 2022**, Nov 2022.
- [n3] **M. Saxon**, S. Levy, X. Wang, A. Albalak, WY. Wang, “Modeling Disclosive Transparency with GPT-2,” **SoCal NLP 2021**, Mar 2021.
- [n2] **M. Saxon**, J. Liss, V. Berisha, “A new model for objective estimation of hypernasality from dysarthric speech,” **Workshop on Signal Analytics for Motor Speech, Motor Speech Conference**, Feb 2020.

Professional Experience

Meta (Facebook Conversational AI) Menlo Park, CA
Research Intern 6/2022–10/2022

Mentors: Chinnadhurai Sankar, Shahin Shayandeh. Through simulated continual learning experiments on publicly available data, we find a decoupling in the catastrophic forgetting exhibited by basic accuracy and the forgetting exhibited by robustness accuracy metrics on dialog state tracking tasks.

Amazon (Alexa Web-based Question Answering) Manhattan Beach, CA
Applied Science Intern 6/2021–9/2021

Mentors: Luca Soldaini, Eric Lind, Rik Koncel-Kedziorski, Alessandro Moschitti. End-to-end spoken QA, multi-modal LM pretraining using mixed phoneme-word synthetic and natural text, AS2 and DPR.

Amazon (Alexa Edge ML) Pittsburgh, PA
Applied Science Intern 1/2020–8/2020

Mentors: Samridhi Choudhary, Joe McKenna, Athanasios Mouchtaris. Investigated the link between semantic complexity of datasets (entropy and graphical measures) and the performance of SOTA E2E SLU models on them, [C6]. Developed a novel SOTA E2E SLU model [C9].

Amazon (Alexa Edge ML) Pittsburgh, PA
Applied Science Intern 5/2019–8/2019

Mentors: Joe McKenna, Samridhi Choudhary, Kai Wei, Athanasios Mouchtaris. Research on neural end-to-end spoken language understanding, runtime early stopping for efficient edge ML.

Aural Analytics Scottsdale, AZ
Research Engineer Intern 12/2018–4/2019

Mentor: Shira Hahn. Development for speech-based clinical neurological health assessment product.

Service

Program Co-Chair, 2022 Southern California NLP Workshop (SoCalNLP) Nov 2022
Reviewer, AACL, EMNLP, ACL, EACL, NeurIPS, ICASSP 2020–present

Mentoring

Mahsa Khoshnoodi, Namrata Mukhija, Fatima Jahara	<i>Fatima Fellowship Mentees</i> , 2023
Avani Tanna	<i>UCSB MS Student</i> , 2022–2023
Andy Ouyang, Daniel Rose, Ryan He, Vaishnavi Himakunthala	<i>UCSB Undergrad Group</i> , 2022–2023
Aditya Sharma, Justin Chang, Nga Ngo, Matthew Ho	<i>UCSB Undergrad Group</i> , 2021–2022
Alex Mei	<i>UCSB Undergrad</i> , 2021–2022
Ayush Tripathi	<i>Visiting ASU Undergrad</i> , Summer 2018

Honors

National Science Foundation Graduate Research Fellowship (NSF GRFP)	2020
University of California, Santa Barbara Center for Responsible Machine Learning Fellowship	2020
Arizona State University Presidential Scholarship (Full Tuition)	2014